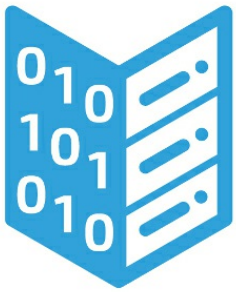

phData

Data Scientist - Project Challenge

Purpose

The goal of this challenge is to give you an opportunity to demonstrate your skills in a reasonably realistic format. In this scenario, you have been assigned a project by a business client and need to complete your analysis and present the results to your client.

Scenario

You have been contracted by a tax firm to help them sell tax preparation software. They have built a dataset over the last 2 years of customer information and recorded whether they were able to sell successfully to each customer. They want to understand this data and build a model to predict if they will be able to successfully sell their software to a given individual. Some of the column names are descriptive, but many are not. The quality of the data cannot be guaranteed, so there may be errors or otherwise dirty data.

Objectives

1. Explore and analyze the data. What can you tell the client about this dataset?
2. Build two models that use the dataset in order to predict the "successful_sell" variable. These two models cannot be the same type of model. For example, if you used logistic regression for the first model, you could not use logistic regression for the second model.
3. Present the results to a business audience and to a technical audience.

Deliverables / Format

The results will be presented to a panel at the beginning of the group interview. The presentation should have two parts. Part one of the presentation should be prepared as if for a client (i.e. not for a group of data scientists) and last 20 minutes. Part two of the presentation should be prepared for a technical audience. This might include a walkthrough of some of the source code, more technical details about algorithms and features, or anything more appropriate for a group of data scientists than a business audience. Part two should also last 20 minutes. After the presentation, we will have Q&A and a general group interview.

You will need to share your presentation deck and source code with us prior to the interview - you can use Google Drive, OneDrive, Dropbox, etc., and share an archive with your interview coordinator.

Scoring criteria

1. Data exploration ability
2. Predictive modeling ability
3. Quality of source code
4. Presentation ability

Keep in mind

- Part one of the presentation should be appropriate for a business client. Imagine someone that is intelligent and skilled at their profession but isn't a data science or mathematics expert.
- In the real world, we would be delivering the source code to the client along with the analysis. The code should reflect good practices.
- We know that you're a busy person! This project is intended to take 4 - 6 hours. It will be judged with that in mind. If there's more that you would do in a real scenario, mention what your next steps would be as part of the presentation. Don't kill yourself trying to put 40 hours of work into this.
- The goal is to demonstrate your competence, communication, and creativity. Have fun with it!